

AL/HR-TP-1994-0021

AD-A284 748



**PRODUCTIVE CAPACITY: THE CONCEPT,
RESEARCH, AND APPLICATIONS**

**Walter C. Borman
Jerry W. Hedge
Paul J. Cook**

**Systems Research and Applications Corporation
2000 Fifteenth Street North
Arlington, VA 22201**

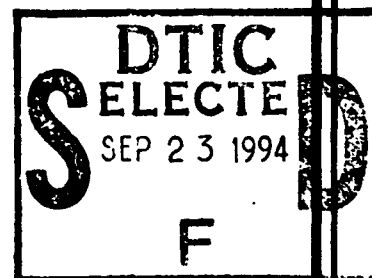
**Donald L. Harville
M. Jacobina Skinner**

**HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352**

August 1994

Final Technical Paper for Period January 1993 - January 1994

Approved for public release; distribution is unlimited.



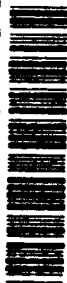
DTIC QUALITY INSPECTED 3

94 9 22 086

**AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS**

ARMSTRONG
LABORATORY

94-30515



copy

NOTICES

This technical paper is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

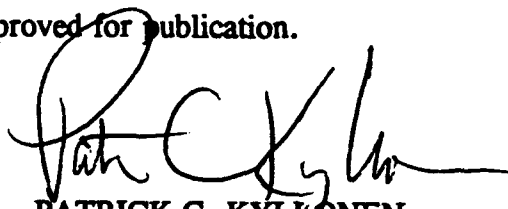
When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.



DONALD L. HARVILLE
Project Scientist



PATRICK C. KYLLONEN
Technical Director
Manpower and Personnel Research Division



WILLARD BEAVERS, Lt Col, USAF
Chief, Manpower and Personnel Research Division

REPORT DOCUMENTATION PAGEForm Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE August 1994	3. REPORT TYPE AND DATES COVERED Final - January 1993 - January 1994	
4. TITLE AND SUBTITLE Productive Capacity: The Concept, Research, and Applications			5. FUNDING NUMBERS C - F49650-92-R-5005 PE - 62205F PR - 7719 TA - 24 WU - 05	
6. AUTHOR(S) Walter C. Borman Jerry W. Hedge Paul J. Cook Donald L. Harville M. Jacobina Skinner				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Systems Research and Applications Corporation 2000 Fifteenth Street North Arlington, VA 22201			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Armstrong Laboratory (AFMC) Human Resources Directorate Manpower and Personnel Research Division 7909 Lindbergh Drive Brooks Air Force Base, TX 78235-5352			10. SPONSORING/MONITORING AGENCY REPORT NUMBER AL/HR-TP-1994-0021	
11. SUPPLEMENTARY NOTES Armstrong Laboratory Technical Monitor: Donald L. Harville, (210) 536-3222				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) This report describes a concept of time-based measurement of the job performance of Air Force enlisted personnel, known as productive capacity (PC). The results of previous PC efforts are reviewed, comments on the theoretical and practical issues that result from this work are provided, and recommendations for future PC research efforts are included. The concept of PC in relation to other criterion concepts and related areas of research is also discussed. The notions of maximal and typical criteria, technical proficiency versus contextual criterion domains, and models of job performance are examined, as well as the literature on time perception/estimation and human learning, in an attempt to identify implications and lessons for thinking about PC. Suggestions are offered for incorporating these notions in future PC research. The final section of this report addresses potential applications of the PC construct in determining manpower requirements and setting enlistment standards. Current practices and possible first steps for implementing PC-based approaches are reviewed. <div style="text-align: right;">DTIC QUALITY INSPECTED 3</div>				
14. SUBJECT TERMS Classification standards Job performance measurements Job proficiency Personnel planning			15. NUMBER OF PAGES 44	
Productive capacity Selection Time-based performance			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

TABLE OF CONTENTS

PREFACE	iv
SUMMARY	1
I. INTRODUCTION	1
II. REVIEW OF PREVIOUS PRODUCTIVE CAPACITY STUDIES	3
III. REFLECTIONS ON PRODUCTIVE CAPACITY	10
Typical Versus Maximal Performance	10
Technical Proficiency Versus Contextual Criterion Domains	14
Models of Job Performance	15
Review of Selected Time Perception and Estimation Literature	17
Human Learning/Skill Acquisition, Retention, Decay/Learning Curve Theory ...	19
IV. PERSPECTIVES ON PC APPLICATIONS	20
Manpower Requirements	21
Standards Setting	22
V. IMPLEMENTATION AND FUTURE PC RESEARCH	26
Thoughts on PC Implementation	28
Summary	29
REFERENCES	30

List of Tables

Table	Page
1. Performance Time Rating Validities	9
2. Speed and Accuracy Correlations	13

List of Figures

Figure	Page
1. Partial Path Model	11
2. Full Path Model	12

PREFACE

This description of a concept for measuring the productive capacity job performance of enlisted personnel is part of an on-going Air Force research program to develop the technology necessary to base selection, classification, and personnel management policies on empirically-derived job performance data. The effort was conducted under Contract F49650-92-R-5005 by Systems Research and Applications Corporation for the Manpower and Personnel Research Division of the Armstrong Laboratory, Human Resources Directorate (AL/HR). The relevant work unit was 77192405, "Improved Methodology for Productive Capacity Measurement."

Accession For	
NTIS CRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

PRODUCTIVE CAPACITY: THE CONCEPT, RESEARCH, AND APPLICATIONS

SUMMARY

Over the last decade the Air Force has been involved in research aimed at identifying, developing, and evaluating job performance measurement technologies. In 1989 the Human Resources Directorate of Armstrong Laboratory began to examine time-based measures of job proficiency, leading to a new stream of research collectively known as productive capacity (PC). This report reviews the results of previous PC efforts, comments on the theoretical and practical issues that result from this work, and provides recommendations for future PC research efforts.

The concept of PC in relation to other criterion concepts and related areas of research is also discussed. The notions of maximal and typical criteria, technical proficiency versus contextual criterion domains, and models of job performance are examined, as well as the literature on time perception/estimation and human learning, in an attempt to identify implications and lessons for thinking about PC. Suggestions are offered for incorporating these notions in future PC research.

The final section of this report addresses potential applications of the PC construct in determining manpower requirements and setting enlistment standards. We review current practices and possible first steps for implementing PC-based approaches.

I. INTRODUCTION

Personnel researchers have long held a keen interest in performance measurement. Virtually all aspects of human resource management rely on some form of performance measures to validate decision tools, such as performance tests, and to provide the basis for personnel decisions, such as hiring, firing, or promoting an individual. Innumerable studies have dissected performance measurement issues from every conceivable angle (e.g., psychometrically, conceptually, legally, practically, econometrically) in a wide variety of work and behavioral contexts. The military has interest in performance measurement for many of the same reasons as business: to support proper personnel decisions and enhance the output and operation of the organization. Productive capacity (PC) is a relatively new approach to performance measurement. We will briefly review other attempts to measure performance in the military to set the stage for our discussion of PC research.

A major stimulus for performance measurement research occurred in the late 70s when the Armed Services Vocational Aptitude Battery (ASVAB) was miscalibrated to its normative base. Many people were accepted for enlistment with lower mental aptitude than as tested (Office of the Assistant Secretary of Defense -- Manpower, Reserve Affairs, and Logistics, 1980), focusing attention on the possible repercussions on individual performance and overall mission capability of the Services. Because there were few methods to measure individual

performance on a wide scale, there were no immediate, reliable answers to the questions about performance and capability. In 1980 the Assistant Secretary of Defense for Manpower, Reserve Affairs, and Logistics directed research be performed to measure military job performance and link enlistment standards directly to performance. Each Service shared in the primary research as well as investing in specific research issues. The Air Force developed a Job Performance Measurement System (JPMS) consisting of hands-on and interview work sample tests, job knowledge tests, and four types of rating forms. Over a 5-year period, instruments were developed and data were collected for eight enlisted Air Force specialties (AFSs) (Hedge & Teachout, 1986; Laue, Hedge, Wall, Pedersen, & Bentley, 1992).

The work sample tests, collectively known as the Walk-Through Performance Test (WTPT), were composed of tasks representative of the job performance domain (20 to 30 tasks for each AFS). Both hands-on and interview formats required the examinee to accomplish the tasks at the work setting under the observation of a trained test administrator, who scored each step in a task as correctly or incorrectly performed.

Four rating forms (task, dimensional, Air Force-wide, and global) were developed to measure job performance from the very specific to the very general for all eight AFSs. Supervisors, peers, and job incumbents made performance ratings on a 5-point, adjectivally-anchored scale. In addition, paper-and-pencil tests of procedural job knowledge were developed and administered to job incumbents in four AFSs.

Correlations between the hands-on and interview work sample tests were moderate to high, ranging from .46 to .84 (median $r = .68$) across the eight AFSs. The correlations between hands-on tests and knowledge tests ranged between .30 and .56 (median $r = .46$). Finally, across self, supervisor, and peer ratings of technical performance (collapsing across the four types of rating forms), correlations with the hands-on work sample tests ranged from .22 to .31, while ratings of interpersonal effectiveness (also collapsing across ratings forms) correlated with the hands-on work sample tests from .03 to .14 (Hedge & Teachout, 1992).

Aptitude and experience relationships with job performance were also examined. The median correlation across the eight AFSs between the Armed Forces Qualification Test (AFQT) score (an ASVAB composite score used by all Services as an indicator of general trainability) and the hands-on performance was .23 (range = .07 to .32). Correlations between total active federal military service (a measure of experience) and hands-on performance ranged between .17 and .38 (median $r = .33$).

While this program of research demonstrated the feasibility of measuring job performance in a military context, the Air Force continued to examine new approaches to job performance measurement. The emphasis in the JPMS had been on the measurement of the *quality* of an individual's job performance. Raters were asked to judge whether incumbents "never meets", "occasionally meets", "meets", "frequently exceeds", or "always exceeds" required standards for quality of work.

In 1989 the Air Force began to examine time-based measures of job performance (Carpenter, Monaco, O'Mara, & Teachout, 1989). Researchers recognized that the JPMS quality-based metric was not ideally suited for enlistment standards setting and that an emphasis on individual differences in the *quantity* of work performed might prove more useful. This decision led to initiation of a new stream of research collectively known as productive capacity, with the expectation that managers might be able to define minimal performance levels more precisely using a quantity of performance scale that reflects worker productivity. Furthermore, measures that quantify individual differences in amount of work performed might be useful for manpower planning purposes. Thus, PC could facilitate the forecasting of aptitude and experience levels needed to accomplish certain quantities of work.

This report examines the PC construct and reviews the background of PC research, the purposes for PC measurement, and the conceptual approaches that drive measurement considerations. A final section addresses potential applications of the PC construct in determining manpower requirements and setting enlistment standards. We review current standard setting practices and propose possible first steps for implementing PC-based approaches.

II. REVIEW OF PREVIOUS PRODUCTIVE CAPACITY STUDIES

This section of the report reviews the PC research program to date. In particular, we review and comment on reports by Carpenter, Monaco, O'Mara, and Teachout (1989); Faneuff, Valentine, Stone, Curry, and Hageman (1990); Skinner, Faneuff, and Demetriades (1991); Leighton, Kageff, Mosher, Gribben, Faneuff, Demetriades, and Skinner (1992); Demetriades and Skinner (1992); Harville and Skinner (1993); and Faneuff (1993).

The technical paper of Carpenter et al. (1989) reported the first attempt to create and investigate a productive capacity index. The approach to measuring PC in this early study was quite different from strategies used subsequently in the research program. Also notable was the attempt to link their PC-related time-to-proficiency (TTP) construct with several of the quality-based JPMS project performance measures.

Carpenter et al. defined PC for a job incumbent as the ratio of estimated fastest possible performance time (in minutes) for a cluster of tasks to observed performance time. They expressed the ratio as $PC = t^*/t$, where t is the observed time and t^* is one minute less than the fastest observed time for all incumbents observed performing a particular task cluster. Note that the values of PC range from 0 for the slowest (least productive) worker to 1 for the fastest (most productive) worker. Conceivably, this ratio scale of measurement allows workers' productivities to be expressed in terms of a proportion of their capability relative to maximum capability. That is, a worker whose $PC = .5$ would be capable of producing 50% of the maximum productivity possible (amount of work per unit of time).

Carpenter et al. gathered data from first term incumbents in the Avionics Communication Specialist AFS. To generate the TTP scores, they first obtained SME estimates of how long the

typical or average airman took to complete tasks in each of 10 task clusters. Then, Carpenter et al. asked each supervisor of the sample participants to identify for *each* of 10 task clusters one of their subordinates whom they believed could accomplish those tasks closest to the average or "normal" amount of time. The chosen subordinate thus became the benchmark performer, and the supervisor was asked to estimate how long it would take each of his/her other subordinates to accomplish the same amount of work as the benchmark performer could do in one hour. This procedure then generated a time for each airman in the sample for that task cluster. The same approach was followed for each of the other task clusters.

These TTP indices were therefore subjective estimates of the relative time (against an average performer) that an airman took to successfully complete tasks. TTP is a quantity-based rather than quality-based index. It has the added advantage that supervisors can probably perform the rating task relatively easily because the comparisons to be made are among airmen whose performance they know well. A problem with the index, addressed in subsequent PC work, is that the comparisons are not made against the same "standard" performer. Average or normal performers may vary considerably across work groups regarding the time it takes them to complete tasks, and thus the comparison airmen (benchmark performers) provide inconsistent stimuli for making TTP estimates for other airmen.

Nonetheless, the correlations between TTP ratings and JPMS performance ratings are instructive. For 58 of the airmen in the total sample, supervisor, peer, and self ratings on the four different JPMS rating forms from the JPMS data collections were available. Although the authors and others (e.g., Faneuff et al., 1990) were not very positive about these results, we see some good evidence of construct validity for the TTP ratings. In particular, correlations between TTP scores and ratings on the more technical proficiency-oriented performance dimensions were quite substantial and, importantly, considerably higher than correlations between TTP ratings and ratings on dimensions less conceptually related to technical proficiency and productive capacity.

For example, focusing on the Global dimension ratings, correlations between TTP and Technical Proficiency were -.52, -.41, and -.50, respectively, for supervisor, peer, and self ratings and only -.19, -.17, and -.06 between TTP and Interpersonal Skills for ratings from the three sources. Similarly, with the Air Force-wide scales, correlations between TTP ratings and the Technical Knowledge/Skill dimension ratings (the most clearly technical proficiency-oriented dimension) were -.55, -.66, and -.46, respectively, for supervisor, peer, and self ratings. Correlations with all other Air Force-wide dimensions were lower. Considering that these correlations were between ratings generated independently by different raters and separated by time, *and* that the TTP ratings reflect an early attempt to measure a PC construct, the research results provide considerable encouragement for continuing efforts to measure PC.

An additional analysis in the report examined the relationship between TTP ratings and: (1) aptitude, derived from ASVAB scores; and (2) experience, defined as number of months in the Air Force. Regression of TTP scores on aptitude and experience scores resulted in an R^2

of .44. The significant beta weights, respectively, were .017 and .027, indicating that experience level was the more important influence on this index of PC.

The Carpenter et al. (1989) predictor model, including aptitude, productivity, cost, and attrition, was unable to accommodate more than one AFS simultaneously. Therefore, Faneuff et al. (1990) used a different methodology to measure PC and addressed its use in setting ASVAB cut scores when several AFSs were considered simultaneously. Here PC was defined as the ratio of t/t^* . Because data were obtained from the JPMS subjects with no direct task performance times available, t was defined as a subject's total WTPT score (sum of hands-on and interview scores) and t^* was the highest obtained total WTPT score for the respective specialties (see the earlier discussion of the JPMS for more details on the WTPT score).

Faneuff et al., showed that aptitude (i.e.; ASVAB) cut scores for an AFS depended on both aptitude standards for the AFS *and* manpower requirements for that and other AFSs. For example, the Carpenter et al. estimate of the optimal minimum AFQT score for the Avionics Communication Specialist AFS was 90. Faneuff et al. argued that this cutoff would be lower if other AFS requirements and the general aptitude level of the recruit pool were considered. Accordingly, they extended the Carpenter et al. model to consider minimum aptitude standards under various scenarios incorporating different recruit pool conditions and manning requirements for other AFSs.

Faneuff et al. recommended more effort toward establishing definitions of optimal performance regarding not only *quantity* (e.g., the t^*) but also *quality* (what they termed q^*). This raises an interesting possibility. The concept of a quantity-based ratio of t^*/t , or the minimum time to complete a piece of work divided by the target airman's time to complete that work, could be extended to quality-based measures. An airman's performance on a dimension q could be compared to the best possible performance on that dimension q^* , and the q^*/q ratio might serve as the quality equivalent to t^*/t . Of course, a quality-based ratio has not been used to date because it is difficult to argue that quality-based ratings can be considered as on a ratio scale.

The Military Testing Association paper by Skinner et al. (1991) reported on the next major step toward measuring PC. The earliest report (Carpenter et al., 1989) recognized that obtaining actual task completion times using hands-on tests was not practical for any large-scale implementation of a PC measurement system. That was the assumption here, as well. Instead, the goal was to get reliable and valid time-to-complete estimates from supervisors. However, the authors also recognized that the comparison rating task used by Carpenter et al., although a good first step toward PC measurement, was conceptually flawed (as discussed previously in this section). This research attempted to develop a rating scale that allowed raters to compare their ratees' time-to-complete with an absolute standard for all raters.

To explore the feasibility of developing a rating scale with absolute standards, Skinner et al. examined a benchmarking idea that would establish actual times to complete a task successfully at three different competency levels -- fastest possible, average or normal, and

slowest possible. The general notion was that if these benchmarks could be reliably determined and placed as anchors on a rating scale, they might serve as useful reference points for raters judging the amount of time *their* airmen ratees spend successfully completing the task.

Accordingly, research proceeded toward developing these anchors for 35-47 tasks from each of four AFSs. Six SMEs from each AFS participated in workshops to establish estimates of these times-to-complete for each task. Specifically, SMEs first made independent estimates of performance times for the fastest possible and the typical or average incumbent *and* for an incumbent who was the slowest possible (but not so slow that the task would be assigned to someone else). Second, the SMEs participated in a modified Nominal Group Technique process that yielded consensus times for each of the three levels. Here are two example tasks with their consensus times:

	Fastest	Normal	Slowest
Replace radio frequency coaxial connector	12 mins.	17 mins.	22 mins.
Perform in-processing of unit personnel	5 mins.	7 mins.	15 mins.

The main finding was that the initial time estimates were reasonably reliable. One-rater reliabilities ranged from .38 to .81 for individual tasks. Also, the fastest, normal, and slowest time estimates had about the same level of interrater reliability. Stepped-up 6-rater reliabilities, appropriate for assessing the reliability of the mean judgments across the six SMEs, were in the .80s and .90s.

A valuable additional benefit of this research is provided by comparisons of the percent time increases for normal to fastest and for slowest to normal in different tasks. As shown in the examples above, for the in-processing task the fastest airman cannot do much better than the normal airman but the normal airman is likely more than twice as fast as the slowest airman. Conversely, the variance across the three levels of the other task is quite uniform. In general, the ratio of the slowest/fastest times *and* the time differences between adjacent anchors provide potentially useful information about tasks. For example, we might predict that more difficult, complex tasks would have larger slowest/fastest time ratios than simpler, more routine tasks. This ratio could in fact provide a numerical index of task complexity.

At any rate, for the four AFSs studied, the slowest to normal percent *decrease* in time varied widely across tasks (6-80%), as did the normal to fastest percent decrease (10-67%). Interestingly though, the *average decrease* for the AFSs were very similar and close to the same for the two comparisons (i.e., slowest-normal and normal-fastest). Those percentages were 39, 37, 37, and 39 for normal to fastest and 32, 31, 32, and 39 for slowest to normal.

The central contribution of this research, however, was a demonstration that performance time anchors for these three competency levels could be reliably estimated by SMEs. The next questions involve actually using these scales to help estimate *t* values for airmen ratees. Can raters (e.g., supervisors or peers) use these scales to provide reliable, accurate, and valid time-

to-complete estimates for ratee task performance? Subsequent research has addressed these questions.

The Leighton et al. (1992) report described data collection for the next logical step in the research program (i.e., are supervisor raters able to make accurate assessments of task performance time [t] for individual airmen?). Three hundred-twenty supervisors in four AFSs estimated time-to-complete (i.e., performance time) for 680 subordinates on multiple tasks (36-50 per AFS). For a subset of the tasks (6-11 per AFS), 240 of the 680 subordinates were actually timed doing hands-on tests associated with those tasks. Their hands-on performance was also rated on a 5-point scale from 1=unacceptable to 5=exceptional. The 3-point was labeled acceptable and could be used as a cutoff below which times would not be counted. This is useful for measuring PC because the definition of t includes the notion of *successful* task completion.

Leighton et al. collected other potentially useful data for studying PC. Supervisors providing the time estimates indicated for each task how often they had observed the subordinate on the task (regularly, occasionally, or never). A fair test of the *potential* validity of the performance time estimates might use only data provided by supervisors who regularly observe the subordinate completing the task. Also, the airmen being tested indicated for each task whether they regularly, occasionally, or never performed the task. Again, subsequent tests of the validity of performance time estimates might consider only tasks that the subject airman regularly performs.

Finally, these researchers administered to members of the subordinate ratee sample, for three of the four AFSs, a job knowledge test, a vocational interest inventory, and a motivation scale. In addition, they gathered a supervisor rating of overall productivity for each ratee. This required the supervisor rater to consider the maximum amount of acceptable work the subordinate can produce in a day and then indicate the percentage of that amount he/she could typically be *expected* to do.

This data collection initiative is a major contribution. First, the time estimation rating comes from supervisors using the improved benchmark performance time scale, with the SME-generated fastest, normal, and slowest times to perform each task included as anchors. Second, the additional measures taken (e.g., job knowledge) can support correlational analyses with the performance time ratings to provide a clearer picture of the PC construct. Third, data on the frequency of observation on the part of raters and frequency of performing the task on the part of ratees might prove very helpful for testing the validity of performance time estimates under relatively ideal conditions. Finally, the sample sizes are reasonably large and subordinate ratees are from more than one AFS, rendering any results from this data set relatively generalizable.

The Demetriades and Skinner (1992) APA paper covered some of the earlier PC results. Demetriades and Skinner showed the interrater reliabilities for the SME benchmarking study. For the four AFSs studied, the authors presented mean 1-rater and 6-rater (the number of SMEs used in the benchmarking effort) reliabilities averaged across all 35-47 tasks, separately for the

fastest, normal, and slowest benchmark estimates. One-rater reliabilities ranged from .48 for the fastest estimates on the Communication and Navigation Systems AFS tasks to .80 for the fastest estimates on the Personnel AFS tasks. Corresponding 6-rater reliabilities ranged from .85 to .96.

In addition, Demetriades and Skinner demonstrated one type of validity for the normal benchmark estimates. A total of 54 airmen were actually timed performing each of 44 tasks (about 11 per AFS of the 35-47 tasks in each AFS), and the mean times taken on each task by these airmen were correlated with the normal benchmark times for the same 44 tasks. The correlation was quite high ($r=.75$, $p < .0001$), indicating that the normal benchmark times on the rating scales mirror rather closely the average time airmen actually take to complete these tasks. It should be made clear (as the authors themselves do) that this finding does not pertain to the validity of performance time ratings of airmen but to the validity or perhaps realism of the normal benchmark times.

Demetriades and Skinner also reported on the correspondence between the mean time taken on tasks by the airmen and the normal benchmark times, assessed in terms of absolute differences between the two. The grand means for the actual times and benchmark estimate times were quite similar (9.20 and 11.68 minutes, respectively, $t=1.32$, not significant). However, Demetriades and Skinner pointed out that the *distributions* of scores across the 44 tasks were very different, with the normal benchmark estimates having much greater variance.

This study's findings suggest that the normal benchmarks for the performance time rating scales are probably realistic, at least in terms of the benchmark times for tasks *relative to* the benchmark times for other tasks. Again, this evidence says nothing about validity related to *performance time ratings of airmen* made on the benchmarked scales, but it is reassuring that the benchmarks supervisor raters will use to make those performance time ratings do not appear to be seriously distorted.

Harville and Skinner (1993) was the first report of performance time rating validities using individual ratees as the level of analysis. Harville and Skinner evaluated the validity and accuracy of supervisory performance time ratings. They defined validity as the correlation between supervisor performance time estimates for individual ratees on a task and the actual time each ratee took to successfully complete that task. Accuracy was defined as the grand means of the performance time ratings across all ratees and tasks for an AFS compared to the grand mean time those ratees actually spend successfully completing the same tasks.

For the analyses, the authors obtained actual time estimates for a moderate sample of airmen in each of four AFSs (up to 61 airmen) on 6 to 11 tasks for each AFS. The airmen who were administered these hands-on tests were also rated by their supervisors on the benchmarked performance time estimate scale for each task. Results are depicted in Table 1.

Validities for three of the four AFSs were moderately positive. With more attention paid to evaluating the validity of performance time ratings on tasks often observed by supervisors *and*

regularly performed by ratees, these validities might improve considerably. We will discuss this topic further in a subsequent section. The accuracy results were promising for two of the AFSs. The results for Avionics Communication would have been more promising, as well, had one of the seven tasks been removed (one of the task times was seriously overestimated).

Table 1. Performance Time Rating Validities

AFS	No. of Tasks	Validity Correlations	Accuracy Results
Personnel	7	.28, $p < .03$	1.44 mins. <i>more</i> for estimated
Avionics Communication	7	.29, $p < .03$	5.01 mins. <i>more</i> for estimated
Aerospace Group Equipment	6	.23, $p < .08$	6.0 mins. <i>more</i> for estimated
Aircrew Life Support	11	-.01, ns.	0.2 mins. <i>less</i> for estimated

Harville and Skinner also investigated the joint effects of experience and ability on the time-to-complete estimates for six of the tasks from each of the four AFSs. Regression of the time-to-complete estimate on experience and ability resulted in Rs from .04 to .31 across the 24 tasks, with a mean of .19. The authors indicated that, generally, experience contributed more than ability to this prediction. It would be enlightening to reestimate the model with actual times rather than estimated times as the dependent variable, although the sample sizes would be much smaller.

Faneuff (1993) used data gathered by Leighton et al. (1992) on the Aerospace Ground Equipment AFS, including performance time rating data for 50 tasks, aptitude scores from the Mechanical composite of the ASVAB, and experience information (i.e., months in the Air Force) on 204 airmen with one to six years in the Air Force. He used the t^*/t formulation, which produces an index that can vary from 0 to 1.0, to compute PC scores for individual airmen on each task. He defined t^* for each task as the fastest performance time estimate for an airman in the sample for that task. The main thrust of his analysis was to regress these PC scores on ability and experience. Faneuff was also concerned about the problem of how to weight PC scores on different tasks to create an overall PC score for each airman. He argued that the most reasonable approach here was to use average time spent data for individual tasks to derive these weights. After several rescaling adjustments for some outlier PC scores, Faneuff conducted the regressions using, essentially, a logistic transformation of the rescaled PC scores.

Resulting R^2 s from the regression analyses for each of the 50 tasks ranged from .01 to .13. Aptitude beta-weights were significantly different from zero for only two of the 50

tasks; for experience, 33 of 50 tasks showed significant beta-weights. In only four of the 50 tasks did a significant aptitude by experience interaction emerge. When PC data were aggregated across the 50 tasks, employing the average task time spent weighting scheme, the resulting R^2 was .16. Experience again had a substantially greater impact than aptitude in prediction of PC. These results confirmed the Carpenter et al. (1989) findings related to the relative weights of aptitude and experience in determining PC. It is noteworthy that the .16 R^2 value was so much smaller than the .44 R^2 derived in Carpenter et al. Possible explanations are that the jobs studied were different, the PC measures were not the same, and the smaller N in the Carpenter et al. analysis may have led to more of an overestimate of R^2 than in the Faneuff study.

Across the studies reviewed above, the PC research program has several features that demonstrate potential for contributing to a practical, useful way of depicting job incumbent and organizational productivity. The idea of the t^*/t ratio for indexing individual incumbent performance on a task, while simple, is a novel way to evaluate performance. That the t^*/t index is useful for thinking about individual performance in a responsive, mission-related way *and* that this index can appropriately address the productivity of units are positive features of the strategy. The conceptualization of productivity in this manner seems both compelling as a new approach in performance measurement and useful for certain important applications.

III. REFLECTIONS ON PRODUCTIVE CAPACITY

In this section, we attempt to provide some perspective on the concept of PC by discussing it in the context of other criterion concepts and related areas of research. The notions of maximal and typical criteria (Cronbach, 1960; Sackett, Zedeck, & Fogli, 1988), technical proficiency versus contextual criterion domains (Borman & Motowidlo, 1993), and models of job performance (e.g., Borman, 1991; Campbell, McCloy, Oppler, & Sager, 1993) are reviewed briefly as they relate to the PC concept. Also, we discuss literature on time perception and estimation, as well as on human learning, in an attempt to identify implications for improving measurement of individual PC.

Typical Versus Maximal Performance

Many years ago Cronbach (1960) made the useful distinction between maximal and typical performance. He referred to maximal, "can-do" performance as ability-related and typical, "will-do" performance as driven more by motivational factors than by ability.

Campbell's (1990) model of the determinants of job performance clarifies the distinction. The model implies that performance is a function of declarative or factual knowledge, procedural knowledge (i.e., knowing how to do a task), and motivation. Maximal performance involves the first two components, related to job knowledge, with motivation essentially held constant. This is because maximal performance measures such as work samples and hands-on performance tests usually constrain workers to try hard for the short duration of the test. Typical

performance, on the other hand, depends substantially on motivation. Will-do, performance-over-time requires job knowledge certainly, but also requires sustained, motivated effort in a setting where motivation is *not* constrained and can clearly vary across job incumbents.

How does ability fit in here? Performance models offered and confirmed by Hunter (1983), Schmidt, Hunter, and Outerbridge (1986), and Borman, White, Pulakos, and Oppler (1991) demonstrate a clear path from ability (i.e., general cognitive ability) to job knowledge to technical proficiency, where the proficiency variable is a maximum performance measure. Figure 1 is a portion of the Borman et al. path model derived from data on more than 4,300 first tour soldiers in nine U.S. Army jobs.

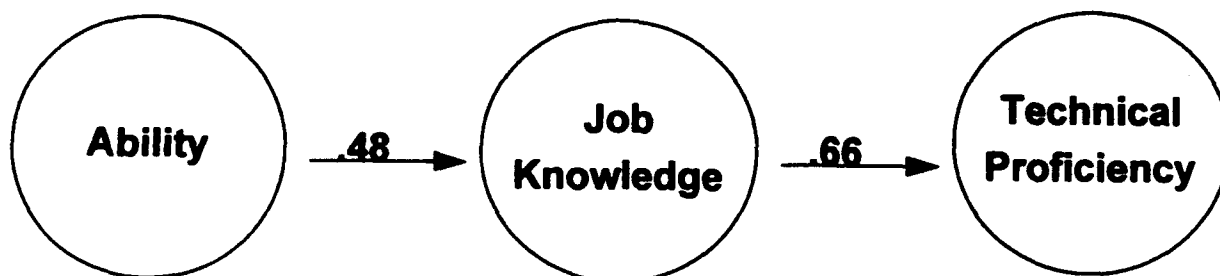


Figure 1. Partial Path Model

Thus, ability appears to influence the acquisition of job knowledge, which in turn impacts upon maximal performance.

Interestingly, when supervisory performance ratings are included in the above model, along with selected personality individual differences variables and behavioral variables reflecting mostly typical performance, the model in Figure 2 emerges (Borman et al., 1991). The fit of this model is very good, with the adjusted goodness of fit index equal to .976 and a root mean square residual of .039.

One way to interpret these results, within the framework of maximal/typical performance and their antecedents, is that the supervisor ratings are likely measures of both maximal and typical performance. Although the ratings are meant to tap typical performance, raters may consider can-do performance when rating on such dimensions as Technical Knowledge and Skill. Accordingly, maximal performance (Technical Proficiency) appears to be a function of ability and job knowledge, but not a function of personality. Typical performance, as captured in the ratings, has as antecedents both the ability→job knowledge→technical proficiency sequence of variables *and* personality through the typical performance behavioral variables. It may be that the technical proficiency ratings path would not be as large if the ratings reflected only typical performance. Although the interpretation of the results is clearly speculative, the performance

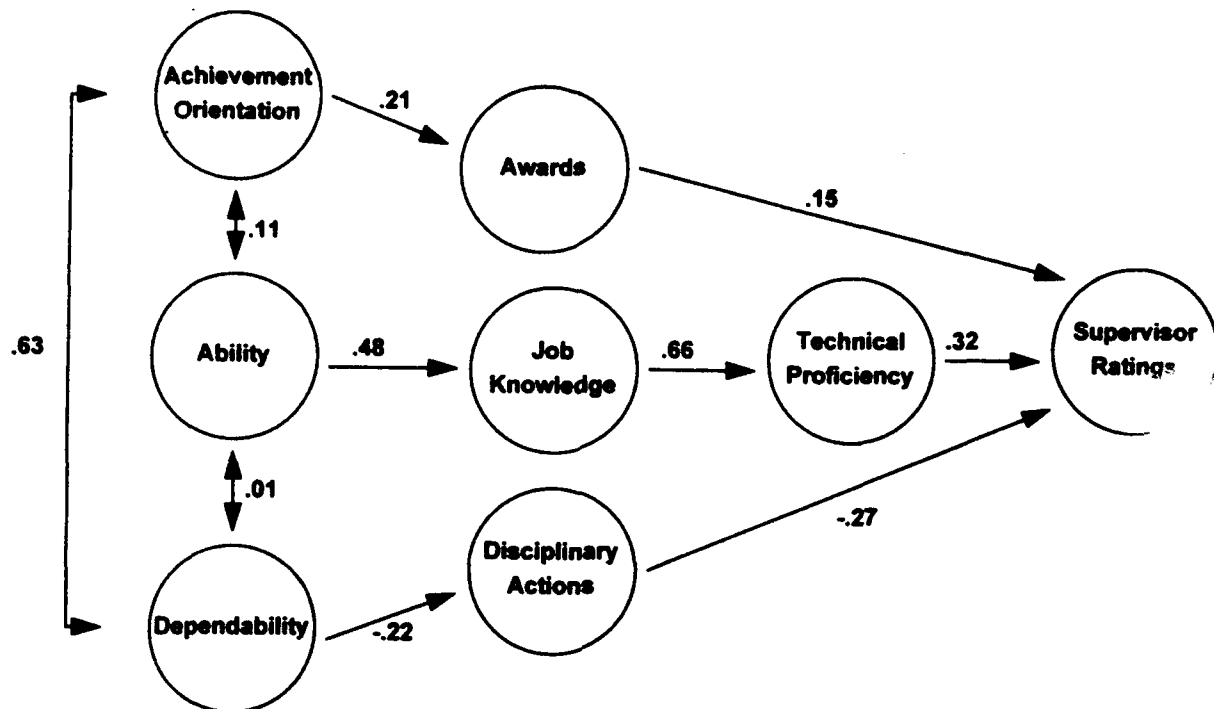


Figure 2. Full Path Model

model suggests that maximal and typical performance can be distinguished by their somewhat different antecedents. Maximal and typical performance, both very important for organizational functioning, may not be very highly correlated.

A study by Sackett, Zedeck, and Fogli (1988) provides a more direct picture of the relationship of maximal and typical performance. Working with grocery clerks, Sackett et al. designed a maximal performance work sample test that consisted of ringing up items in a standardized shopping cart. Measures of speed (total time to complete the cart) and accuracy (number of voids and incorrect entries) were derived from this test. For the typical performance measure, a computer-monitored system kept track of each grocery clerk's performance over a 30-day period. Speed (number of items rung up per minute) and accuracy (number of voids per day) were measured. Sackett et al. did not differentiate clerks based on type of register (scanner or electronic) used; a later study by DuBois, Sackett, Zedeck, & Fogli (1993) did so and those data are shown in Table 2, for more 550 grocery clerks on a scanner type register and for more than 135 clerks on an electronic type register.

The correlations between the typical and maximum measures of speed were .32 ($p < .01$) on the scanner type register and .11 on the electronic register, and between typical and

Table 2. Speed and Accuracy Correlations

	TS	MS	TA	MA
Scanner register				
Typical Speed	.---			
Maximal Speed	.32**	.---		
Typical Accuracy	-.04	.02	.---	
Maximal Accuracy	.06	.14*	.12**	.---
Electronic register				
Typical Speed	.---			
Maximal Speed	.11	.---		
Typical Accuracy	.02	.24*	.---	
Maximal Accuracy	.20*	.09	.32**	.---

Note. TS = typical speed; MS = maximal speed; TA = typical accuracy; MA = maximal accuracy.

* $p < .05$, ** $p < .01$

maximum measures of accuracy were .12 ($p < .01$) on the scanner register and .32 ($p < .01$) on the electronic register. Thus, three of the four comparisons for linking maximal and typical performance in the two samples of clerks yielded significant results, but the most striking finding here is the generally low correlations between the two components of performance.

In addition, DuBois et al. (1993) demonstrated a somewhat different pattern of antecedents or predictors of maximal and typical performance. In particular, for the scanner sample, a test of perceptual ability predicted maximal performance (speed) significantly better than it predicted typical performance related to speed. Similarly, the numerical ability-maximal performance (speed) correlation was significantly higher than the numerical ability-typical performance correlation for speed in the electronic sample. Again, a more substantial link exists between ability-related predictors and maximal performance compared to ability-typical performance relations.

So, what does all of this have to do with PC? We see two important implications of the above work for the PC research program. First, PC is almost certainly a maximal performance

concept. The intent is to measure *capacity* for performance rather than typical levels of performance-over-time. However, it should be recognized that the PC construct may be deficient as a criterion in the sense that it does not address the typical performance domain. This point would be moot if the literature indicated high correlations between maximal and typical performance, but, as we have seen, relationships between these constructs are not very strong (Sackett et al., 1988).

Second, ability should be the most successful individual differences predictor of maximal performance. As mentioned, DuBois et al. (1993) found that their ability measures correlated more highly with maximal performance than with typical performance. Accordingly, it will be important to continue including ASVAB variables in the research on PC. General cognitive ability and related abilities and aptitudes *should* predict PC.

Technical Proficiency Versus Contextual Criterion Domains

Another distinction related to performance constructs in the criterion space is between task performance and contextual performance. Borman and Motowidlo (1993) recently discussed this distinction and its implications for personnel selection programs. Task performance relates to the proficiency with which job incumbents perform activities that are formally recognized as part of their jobs. They defined contextual performance as going beyond the activities that comprise "the job" to help accomplish organizational goals. Integrating elements of organizational citizenship (e.g., Organ, 1988), prosocial organizational behavior (e.g., Brief & Motowidlo, 1986), and a model of soldier effectiveness (Borman, Motowidlo, & Hanser, 1983), Borman and Motowidlo identified five contextual dimensions. They are listed below.

1. Persisting with enthusiasm and extra effort as necessary to complete own task activities successfully
2. Volunteering to carry out task activities that are not formally part of own job
3. Helping and cooperating with others
4. Following organizational rules and procedures
5. Endorsing, supporting, and defending organizational objectives.

There are conceptual, and at least limited, empirical arguments that suggest contextual performance is related to organizational effectiveness. That is, organizations with individuals who are effective in the contextual performance domain will tend to be more effective than organizations whose incumbents are ineffective in this domain (Borman & Motowidlo, 1993). Therefore, contextual performance can be seen as important for successful organizational functioning.

In fact, a recent study by Motowidlo and Van Scotter (in press) sheds light on the relationships between task performance and contextual performance in an Air Force enlisted sample. Motowidlo and Van Scotter found that both task and contextual performance related

substantially to overall performance (correlations of .46 and .41 respectively);¹ however, the two types of performance correlated only .17 with each other. Thus, this study provides more evidence for the importance of contextual performance in contributing to overall effectiveness, *and* the low correlation with task performance indicates that contextual performance cannot be "covered" by measuring technical proficiency only.

Our point here is that PC is measuring the task performance, technical proficiency criterion domain and is *not* tapping the contextual performance domain. Parallel to the maximal-typical performance discussion, this means that we must acknowledge that PC measures by themselves will be deficient as criteria. Also parallel to the previous discussion, ability predictors are most likely to be correlated with task performance, arguing again for including ASVAB variables in the PC research program.

Another study used the Air Force JPMS database to investigate the link between ability and technical proficiency. Alley and Teachout (1990) examined the joint effects of aptitude and experience on job performance. Job performance was measured using the hands-on work sample tests. Aptitude was measured using scores on the ASVAB. Experience was defined for each job incumbent as their total active federal military service at the time of testing. Regression analyses were employed to examine the unique contributions of aptitude and experience to the prediction of job performance, and to assess the degree of interaction between the two variables.

Alley and Teachout (1990) found that both aptitude and experience made separate and unique contributions to prediction of performance. Persons with higher aptitudes performed better than those with lower aptitude, while individuals with more experience performed better than those with less experience. Interaction effects were significant in only one of the eight AFSs. In addition, curvilinear effects were discovered with experience in one AFS, with aptitude in three AFSs, and with both aptitude and experience in one AFS. As noted by the authors, the absence of significant interaction effects is consistent with the previous work of Schmidt et al. (1986, 1988). An important message is that, again, ability and aptitude are substantially related to technical proficiency performance criteria.

Models of Job Performance

To continue the discussion of PC in the context of other criterion concepts, recall that we discussed performance models that included ability, job knowledge, technical proficiency, and ratings (Hunter, 1983; Schmidt et al., 1986). Also described was an expanded model that added personality and selected behavioral variables (Borman et al., 1991). Results of these efforts were used to make inferences about the likely antecedents of maximal and typical performance.

¹ Importantly, Motowidlo and Van Scotter had *different raters* evaluating task, contextual, and overall performance.

In this section, we describe an attempt to identify latent performance constructs in common across all first tour U.S. Army soldiers (Campbell, McHenry, & Wise, 1990), and even more ambitiously, an attempt to explicate a performance model to fit all jobs in the Dictionary of Occupational Titles (Campbell, McCloy, Oppler, & Sager, 1993). We do this in order to provide more perspective on what parts of the criterion space PC is likely to measure and what parts are likely not covered.

Campbell et al. (1990) developed latent variable models to explore the structure of job performance for first tour soldiers in the U.S. Army. Confirmatory factor analyses of correlations between a wide array of performance measures suggested a 5-factor model. Those factors are:

1. Core Technical Proficiency
2. General Soldiering Proficiency
3. Effort and Leadership
4. Personal Discipline
5. Military Appearance and Physical Fitness

The first two factors are dominated by hands-on and job knowledge test scores. Factors 3 through 5 are defined largely by performance ratings in these criterion domains and certain administrative measures such as number of awards/commendations and number of disciplinary problems. It appears that PC is closely aligned with the first two factors. However, the more volitional, extra-technical proficiency factors are not covered by the PC concept.

More recently, Campbell et al. (1993) introduced an 8-factor model that purports to reflect, at the most general level, all performance requirements for all jobs. Not all of the factors are relevant for all jobs, but the taxonomy is meant to be exhaustive of all possible performance requirements. The factors are listed below:

1. Job-Specific Task Proficiency
2. Non-Job-Specific Task Proficiency
3. Written and Oral Communication Task Proficiency
4. Demonstrating Effort
5. Maintaining Personal Discipline
6. Facilitating Peer and Team Performance
7. Supervisor/Leadership
8. Management/Administration

Again, we see that PC is likely to cover only some of this "population" of dimensions. PC seems relevant for Factors 1 through 3 but not so relevant for Factors 4 through 8.

Review of Selected Time Perception and Estimation Literature

Outside of the work done with time-based performance measurement by the PC research program, few studies can be found that provide useful information to support or argue against such a system. A brief review of the time perception and time estimation literature offers some tangential, but relevant research findings to consider for further PC study.

Two opposing theories of time perception are discussed in the literature, one by Ornstein (1970) and one by Priestly (1968). Ornstein's theory is based on memory storage size and asserts that time estimates are a function of the amount of effort expended for information processing. Thus, the more complex the stimulus is, the more processing it will require, and the time period in which processing occurs will seem longer. Priestly's theory, on the other hand, is based on the adage "time flies when you're having fun," meaning that the more actively one must process stimuli, the faster time seems to pass. Thus, time estimates according to Priestly are an inverse function of stimulus complexity because the higher information processing requirements demand the cognitive attention that would otherwise be devoted to time estimation. Ornstein's theory would seem to support the findings of Harville and Skinner (1993) that time estimates were more accurate for shorter duration tasks which, presumably, present less complex stimuli.

Hartley, Brecht, Pagerey, Weeks, Chapanis and Hoecker (1977) focused on self-reports by workers of task identification, rank ordering of tasks according to time spent, and actual time estimation per task. Results showed that the accuracy of self-report time estimates decreased with the increase in measurement scale complexity. In other words, the time estimation accuracy suffered as responses progressed from nominal (task identification) to ordinal (rank ordering) to ratio (time estimation). Since accuracy was poorest at the ratio level where subjects were asked for specific time estimates, the authors suggested that objective observers be used in situations requiring precise task time measurement. The Hartley et al. (1977) study also supports the short duration task accuracy claims of Harville and Skinner (1993), but they suggest using actual timed performance whenever possible.

Carroll and Taylor (1969) compared the time allocation estimates of clerical workers with time measurements obtained through unobtrusive work sampling over a two-week period. They found that the average correlation between estimated and actual time allocations was .88, a finding supportive of the notion that at least gross estimates of relative time spent are accurate. These authors maintained that low-level employees can easily and accurately provide time estimates to serve as a general guide to the type and frequency of tasks performed on various jobs, which would be useful for numerous personnel functions.

Along these same lines, Christal and Weismuller (1988) recommended that job descriptions be based on the percentage of work time that workers spend on each task. They note, however, that personal experience indicates that many workers do not have a clear idea of the exact percentage of time devoted to each task that they perform. This experience is corroborated by other research (e.g., Carpenter, Giorgia, & McFarland, 1975; Wilson &

Harvey, 1990). As a result, they encourage the use of a "relative time spent" scale, believing that people can state with confidence that they spend more time on one task than on another.

Turney and Cohen (1978) examined the estimation of effort expended by asking information processing personnel to rate the duration of task performance. Despite this somewhat different orientation, their dependent measures were similar to PC work, namely perceived time and monitored time. The former was obtained through ratings of effort and time spent on three tasks by Army personnel and the latter was measured directly by a computer over seven weeks. The perceived and monitored time measurements were highly consistent, and effort and time were significantly and positively correlated. However, the strength of the relationship varied across tasks and time measurement sources.

Another study (Troutwine, 1984) examined the related issue of perceived task quality and time estimates. Subjects in this study listened to two audio tapes, one containing a boring ethics passage and the other a tale of mythological adventure, and then rated the tapes according to how "interesting" and "pleasant" each was. They also estimated the time interval for each tape. Troutwine found that as the tape's favorability rating increased, the estimated duration decreased.

Frederickson (1988) examined a theory of temporal experience that states that experience of time is an interrelation of two components, succession and duration. Succession is also known as temporal orientation; one can have a future, present, or past orientation to time. Duration is also called temporal pace, which is the amount of time elapsed in an interval. Frederickson studied the impact of temporal orientation on *temporal pace* by analyzing the verb tense used by clients in a 40 minute therapy session (orientation measure) and obtaining a written estimate from each client of the time elapsed during the session (pace measure). Results showed that those clients with a "past" orientation had significantly longer time estimates of the session (slow pace) than did clients with a present or future orientation. Frederickson (1988) suggested that both components be assessed simultaneously when considering the influence of temporal experience on human behavior.

Hogan (1978) proposed a time estimation theory that considers the combined contributions of stimulus complexity with the personality dimension of extroversion. He explained that extroverts have a higher stimulation baseline than introverts and thus will perceive a simple, "boring" stimulus as taking up a longer time period than an introvert would estimate. Hogan argued that Ornstein and Priestly are both right, but that the true relationship between stimulus complexity and time estimation is not linear but curvilinear. When considering individual information processing styles in the form of extroversion, an inverted U-shape best represents the predictive relationship. Specifically, there is an optimal moderate amount of stimulus complexity (that is higher for extroverts than introverts) above or below which information processing slows and estimates of time duration increase. Hogan suggested that whether one is understimulated or overstimulated, either state is boring and thus perceived to be lengthy.

While Hogan provided no empirical evidence, Zakay, Lomranz and Kaziniz (1984) tested his theory, and found support for Hogan's hypotheses. However, stimuli used in this study were slides of geometrical figures and complexity was determined by the number of internal angles in the figure. Also, time estimates all averaged under 10 seconds, so applicability of the Zakay et al. (1984) findings to productive capacity time estimates is tenuous.

These time estimation and perception studies provide a variety of interesting, and tangentially related information. Collectively, these studies suggest the need for rigor in PC research design because so many content and context variables may influence the accuracy of time estimation.

Human Learning/Skill Acquisition, Retention, Decay/Learning Curve Theory

The human learning area, with its voluminous body of literature, is well beyond the scope of this paper. Much has been written about even small subsets of the human learning domain. Still, a few general comments and reflections about these issues and their relation to the PC research program are appropriate here.

People differ considerably in the skills they achieve for complex tasks, even after extensive training. In addition, prolonged practice or experience with some tasks may even increase individual differences (Fleishman & Mumford, 1989). Cognitive psychology and information processing approaches to learning discuss the link between individual differences and task performance. Fitts and Posner (1967) noted three phases of skilled performance during skill acquisition: cognitive (declarative knowledge), associative (knowledge compilation), and autonomous (procedural knowledge). Shiffrin and Schneider (1977) argued that not all tasks allow skill acquisition along the stages described by Fitts and Posner. Rather, tasks that require learners to deal with novel situations/demands never allow them to progress beyond the first or second stage. Shiffrin and Schneider (1977) described these novel or inconsistent tasks as requiring "controlled" information processing, while simple, consistent tasks allow "automatic" information processing.

Ackerman (1986, 1987) and Ackerman and Humphreys (1990) proposed that three major ability classes relate to individual differences at these three stages of learning. During Phase 1 general intelligence is critical; during Phase 2 knowledge compilation involves integration of the cognitive and motor processes required for performing a task, while Phase 3 occurs when the individual has automatized the skill. These phases of skill acquisition are consistent with the notion of three skill acquisition stages: (1) novice, (2) journeyman, and (3) master.

Unfortunately these learning stages often interact with, and are affected by, the military training/performance environment. By its very nature, the military environment fosters/requires (1) high levels of planned and unplanned turnover, (2) wide variation in task content and task difficulty, and (3) extreme variability of initial skills and ability in the entrant population (Lane, 1987).

The regular cycle of change and progression within the military environment means that as the airman begins to move toward the "mastery" stage, and even the "journeyman" stage, as often as not he/she begins to assume supervisory responsibilities, and the frequency and recency of technical task performance may begin to decline.

Not much is known about how an individual trainee's characteristics are related to long-term retention of skills (Fleishman & Mumford, 1989). However, a study by Fleishman and Parker (1962) suggests that levels of retention/decay were explainable in terms of individual differences among subjects in the habits acquired during practice of the original task. In other words, final level of proficiency during training was an important factor in level of proficiency maintained.

These issues of skill acquisition and decay may play an important role in the PC measurement research. Of special note here are the findings by Faneuff (1993) that PC initially increases with experience until it reaches a maximum and then begins to steadily drop off, as shown in his plotting of response curves. Faneuff noted that the decreasing PC with increasing experience over a portion of the curves and surfaces was unexpected. He speculated that there might be some point in an airman's career where he or she may begin to experience skill degradation.

Throughout the literature there are regularities in the general form of learning and acquisition curves. As noted by Lane (1987), the negatively accelerated curve is not only "typical" of group performance, but it is found in most real-world training situations. In addition, the general shape of the learning curve is consistent with all major theoretical explanations of how skill acquisition proceeds. However, Lane also noted that while the negatively accelerated shape is quite common, parameters of the curves tend to be situation and task dependent. Thus, the particular curve "family" providing the best fit to a given set of data is likely to vary as a function of the task, its components and difficulty level, the characteristics of the people performing the task, the length of the practice (as well as the frequency and recency of performance), the way performance is measured, and the training method used.

What does all of this mean for the productive capacity research project? Performance is a reflection of these acquisition, retention, and decay patterns. Performance is also linked to aptitude, ability, and task characteristics. Consequently, it becomes crucial for PC research to gather as much data specifically related to these variables as possible.

IV. PERSPECTIVES ON PC APPLICATIONS

We see two related applications stimulating future research on PC: determining manpower requirements and setting standards to achieve those requirements. Manpower and personnel are two separate functions in the Air Force. Manpower planners determine the number and skill levels of positions required to perform specific jobs at locations worldwide. Personnel specialists select, classify, and allocate people to the manpower positions.

What is missing in this process is attention to the broad range of capabilities possessed by individuals. We believe PC, with its interactive effects of aptitude and experience, can contribute to better management of manpower and personnel functions within the Air Force.

Manpower Requirements

The Logistics Composite Model (LCOM), created in the late 1960s, is the accepted Air Force manpower model (Boyle, 1990). This model is a policy analysis tool which relates base-level logistics resources with each other and with sortie generation capacity. The logistics resources modeled in LCOM include maintenance personnel, spare parts, and aerospace ground equipment. When using this model, the spare parts are constrained, then manpower constraining is performed. With the objective of maximizing sortie generation potential, LCOM is used to prevent manpower staffing from being too high (i.e., idle or underutilized) or too low (i.e., too busy or overutilized). The manpower for each AFS is optimally constrained, when adding manpower does not affect sortie rate, and reducing manpower would drop the sortie rate below the desired level.

Due to the lengthy process involved in determining LCOM manpower estimates, the Queuing Manpower Model (QMAN) was developed (Grobman, Quick, & Weaver, 1994). This model applies a queuing algorithm to AFS/crew size clusters to determine the manpower necessary to meet flying demands. This value is then compared to utilization and crew size effects to determine the actual manpower requirements. QMAN provides rapid manpower estimations that correlate well with the estimates from LCOM. These faster estimations give QMAN capabilities, such as determining how specialty structuring affects manpower requirements, that LCOM does not have. Also, QMAN has the PC relevant capacity to demonstrate that as the task performance times of aircraft maintainers decrease, the number of aircraft maintainers that are necessary to support flying decreases.

With models such as QMAN available, time-based, quantitative measures would be useful to Air Force manpower planners. Currently, the aggregate of all manpower positions, tempered by total mission objectives for a given year, forms the basis for budget submissions to sustain the objective force. Congressional authorizations in turn specify the number of people the Air Force can expect to fill its manpower requirements. Requested manpower and authorized end strength seldom match precisely.

Once end strength is set, personnel planners determine the number of accessions required to fill projected vacancies. They also determine the training requirements for new people. Air Force applicants are selected for service based on measures of individual differences in aptitudes, assuming that some minimum degree of mental quality is necessary to successfully accomplish entry level jobs. This process also implicitly assumes that, once trained, airmen will achieve desired proficiency sometime before expiration of their enlistment and that performance will increase with experience to prepare them for higher skill activities. However, newly trained airmen may be able to meet job quality requirements, but the speed or quantity of their performance on the job may differ substantially from the expected level.

PC has potential to enhance manpower planning by considering tradeoffs in amount of work expected in given periods of time by people who have different qualifications. For entry-level manpower planning, PC offers tradeoffs between the number of people required and the quality of people (fewer high-quality people or more low-quality people). Such tradeoffs could be worked into the endstrength plan and made into recruiting objectives. Personnel planners could prioritize personnel quality goals according to importance of jobs, or job locations, for specific mission needs.

For manpower planning to sustain a specified force, experience could be factored in, to allow tradeoffs with the number of people required (fewer people who have extensive experience or more people who have limited experience). Personnel planners in this context could establish cross-training or reenlistment objectives to achieve the desired force capability.

If research demonstrates consistent relationships of PC with aptitude and experience in a variety of job specialties, the ratio properties of PC could form the basis for quantity/quality and quantity/experience tradeoffs discussed above. Conceivably, two airmen with $PC = .5$ can produce the same amount of work as one airman with $PC = 1.0$ or four airmen with $PC = .25$. By considering the design of specific job tasks (independent or team performance, backup requirements, etc.), recruiting budget (realistic quality goals), and endstrength limitations, planners could develop manpower requirements to maximize expected performance. Research is needed to develop models that will take all of these factors into consideration.

As an adjunct to the manpower planning process, redesign of individual jobs themselves might be enhanced by study of PC. Because PC measurement addresses time to perform specific tasks, it may be possible to group tasks in more logical ways to enhance performance. Regrouping tasks may then suggest more logical organization of jobs themselves. Also, consideration of different requirements for jobs under peacetime or wartime conditions may affect job design.

Standards Setting

The perspectives on PC application for determining manpower requirements are consistent with the direction of past research and point to ways to fill gaps in personnel management within the Air Force. Likewise, job selection standards based on prediction of PC can potentially aid in selecting people capable of performing at desired levels within time objectives. Much of today's standards setting process is judgmental -- commanders, trainers, and functional managers assess the capabilities of assigned personnel and adjust standards to modify capabilities. PC offers a more sophisticated, empirically based source of information to help make these judgments. Although these directions appear fruitful, there are many specifics to be developed. The discussion here is intended to lay out current thinking and draw upon previous studies that may support future PC research. In this section we first define the problems that standards setting addresses. Second, the major methods of setting standards are described. And third, we speculate on how this concept may be relevant to PC-based standards.

Performance is typically measured on a continuum (e.g., from very effective to very ineffective performance). Furthermore, most researchers concerned with performance measurement would likely judge such measures to be on an ordinal scale, with performance scores comparable only in terms of rank order (for example, many rating scales use terms like "outstanding" or "average" which do not convey information about absolute amount of performance or amount of performance more than the next lower level). It is possible that with some performance measures, interval scale assumptions may be defensible. For example, with hands-on, work sample performance tests, the percent correct scores may be argued to possess interval properties (i.e., the number of units between two levels of work can be determined). Whether ordinal or interval assumptions hold, continuous scales for indexing performance are useful, as long as we are simply comparing performance levels among incumbents.

There are applications, however, where we would like to measure performance dichotomously. An important practical question to ask in some situations may be, "Is this incumbent qualified or not?" or "Is he or she a satisfactory performer or performing at less than an acceptable level?" Dichotomous performance measurement may be useful in at least three human resource applications. First, the identification of training needs can benefit from knowing whether individual incumbents are performing satisfactorily or unsatisfactorily. A train/do not train decision for an incumbent is simplified if we know that the person is qualified or unqualified in a particular aspect of the job.

Second, the effect of other personnel programs or interventions can be meaningfully evaluated and relatively easily explained to managers by comparing the numbers or percentages of incumbents performing satisfactorily after the program or intervention with the numbers or percentages performing at a satisfactory level before the program or intervention. Third, goal setting is most effective when specific, non-ambiguous goals are set (e.g., Locke, Shaw, Saari, & Latham, 1981), and therefore a goal of becoming qualified on a task or job, where the standards for qualification or satisfactory performance are well specified, is likely to provide the desired effort and motivation outcomes.

Accordingly, for these and related human resources applications, standards setting, with the accompanying capability to measure performance dichotomously, satisfactory or unsatisfactory, qualified or not, has some definite advantages. We now briefly review methods for establishing such standards. Advantages and disadvantages of each method using PC are also noted.

Broadly speaking, there are two primary approaches for setting performance standards in the context of mental abilities/aptitude testing: item-based and examinee-based methods. Item-based methods focus on the performance test and evaluate how a minimally competent person is likely to score on the test. Examinee-based methods start with identifying competent and incompetent incumbents and then derive a performance test cutoff score separating the two groups.

Item-Based Methods. Nedelsky (1954) espoused a method for determining minimum acceptable scores on multiple choice paper-and-pencil tests. His approach involves having each job expert (experienced job incumbent or well-qualified supervisor) review each test item and identify response alternatives that the minimally competent incumbent would *not* pick as a correct answer. Then, for that expert, the reciprocals of the number of response alternatives for each item not so identified are added to obtain a minimum passing score. Thus, for a hypothetical 2-item test, with 4 response options for each item, suppose an expert identified 2 response options with each item that the minimally competent examinee would *not* pick as correct. The minimum acceptable score then would be $1/2 + 1/2 = 1$. The actual cutoff score for a test is established by averaging these scores across several experts.

The Angoff (1971) method requires a somewhat different judgment on the part of job experts. Each expert is to consider a *group* of minimally competent workers and estimate the percent of this group that would get each item on the test correct. These percentages are then averaged across all items on the test and, as desired, across experts.

Ebel's (1972) strategy for determining a minimally acceptable score on a test is quite a bit more complex. To employ this method, a 2-dimensional matrix of item difficulty (e.g., easy, medium, hard) by relevance of item for successful performance (e.g., essential, important, questionable) is first constructed. Then, each job expert sorts each test item into one of the cells and answers the question for each cell, "If a borderline examinee were to respond to items like these, what percentage of the items would be answered correctly?" These percent estimates are subsequently averaged, weighted by the number of items in each cell. Again, typically, results from several experts are averaged.

Jaeger's (1982) method is more straightforward. Job experts are asked to review each test item and answer the question, "Should *every* examinee who is at least minimally competent answer this question correctly?" The number of items for which the answer is "yes" then becomes the minimally acceptable score, and these scores are usually averaged over experts.

Unfortunately, each of these methods has substantial problems with respect to its relevance to PC. There are many criticisms of Nedelsky's (1954) method, including low confidence levels reported by experts making the judgments (Poggio, 1984) and unrealistic assumptions about the likely judgment process experts employ in identifying response options (Jaeger & McNulty, 1986). However, the overriding practical problem is that the method can only be used for multiple choice tests.

The Angoff (1971) strategy is more promising for PC-based standards setting. Psychometric properties of the expert judgment resulting from the method are reasonably good (Norcini, Lipner, & Langdon, 1987). Application of the method would, however, seem to require a somewhat different judgment task on the part of the job experts. Perhaps experts could directly estimate the time to completion for a minimally qualified and minimally acceptable incumbent on each task. Experts might be aided by the fastest, typical, and slowest benchmark times for each task, provided they had been previously estimated. One would suppose that the

"correct" estimate for this judgment would be somewhere between the typical and slowest times. Similar to what was done for the benchmark estimates, interrater agreement between experts might be used to evaluate the quality of the minimally qualified time estimates.

Ebel's (1972) method does not appear applicable to PC measurement. Tasks might be treated as items and sorted into a difficulty by relevance matrix, but asking experts to estimate the percent of tasks minimally competent incumbents could get "correct" (e.g., less than the normal time?) does not seem very useful.

Finally, Jaeger's (1982) approach could perhaps be adapted to develop minimum acceptable time estimates for tasks. Instead of asking job experts to make direct time estimates for minimal competence on each task (as was suggested previously), they might be instructed to identify for a task the longest time to completion that "every examinee who is at least minimally competent" would be allowed.

Examinee-Based Methods. Zieky and Livingston (1977) developed what they called the borderline-group procedure. With this method, job experts knowledgeable about the performance of several incumbents are asked to sort these incumbents into three groups -- competent, borderline, and incompetent. The median test score for the borderline group is then incorporated as the minimally acceptable score.

The contrasting-group method is similar to the borderline-group procedure. Livingston and Zieky (1982) suggested asking job experts to identify incumbents who are definitely qualified and incumbents who are definitely not qualified. Then, performance test scores for the two groups can be plotted, and the minimally acceptable score is set at the point where the two distributions of scores intersect.

Both advantages and disadvantages of these two methods have been discussed in the literature. Example advantages are the relative simplicity of the procedures and the fact that they can be used with any kind of performance test or rating (see Poggio, 1984). Disadvantages include the requirement that the sample of job incumbents used in these procedures must be representative of the target population of incumbents and, with the borderline-group method, identification of the borderline group (usually a very small number) may be difficult (Jaeger & McNulty, 1986).

With respect to PC-based standards, it is possible that these examinee-based methods could be used to help evaluate the validity of job expert minimally acceptable time estimates, for at least a few tasks. After a minimal competency time has been estimated for a task, job incumbents might be identified as competent or incompetent (and perhaps borderline), and actual time to completion scores for hands-on performance on the task assessed for each incumbent. Using the contrasting-group method, for example, the intersection of the time to complete score distributions could then be noted, and that time compared to the time estimate made using the revised Angoff or Jaeger methods. Similar to Skinner, Faneuff, and Demetriades (1991), correlations between the two sets of estimates might be computed, provided that a reasonably

large number of tasks were included. This correlation would then serve as one index of the validity of the minimally acceptable time estimates.

Our brief review of the literature suggests several ideas applicable to PC-based standards setting. Both the item-based and examinee-based approaches to setting performance standards offer methodologies that could prove useful. Either a modified version of the Angoff or Jaeger approaches to item-based standards setting could be adapted for PC measurements. The Angoff method would require experts to directly estimate time to completion per task for a minimally qualified incumbent. Jaeger's approach would focus on having experts identify for a task the longest time to completion allowable for a minimally competent examinee. The examinee-based methods may be more useful for validating the item-based procedures than for setting performance standards. Correlations between contrasting-group (or borderline-group) scores and scores derived from item-based estimates could be used as one index of these time estimates.

V. IMPLEMENTATION AND FUTURE PC RESEARCH

Earlier we raised a few issues relevant to the application of PC research to establishing manpower requirements and setting standards. We then discussed how some of the current standards setting methods relate to PC. Here we return to the broader framework for considering PC within the context of setting standards to maximize military output for a given cost. All new ideas for better personnel management ultimately face the test of cost effectiveness -- does the idea provide something of value that justifies an investment in it? Although PC research is far from generating operational systems, the research should proceed with implementation in mind. In this section we will turn to thoughts developed by Black (1988) as he considered the DoD Job Performance Measurement Project and looked ahead to implementing performance-based measures for setting standards. The issues he raised are equally relevant to PC.

Black's chief concern centered on the impact on overall military capability from new standards setting approaches that seek to maximize individual performance. If an objective of new standards is to allow tradeoffs between greater military output associated with higher-quality entrants and their higher cost compared with lower-quality entrants, seeking maximum individual performance alone may result in suboptimal quality solutions for the military overall. Such objectives tend to miss the impact of quality on unit or group output, which is more closely aligned with military capability. Individual performance has an interactive effect on group output (rather than a simple, unadjusted summation of independent efforts). Black raised a number of issues to consider in research on performance-based standards setting. In summary, these issues are:

- The military personnel systems are closed systems. An individual typically enters at a low level (unskilled and low ranking) and the Services invest heavily in training and career development as the person progresses upward. Recruiting and training programs are expensive. Services have numerous options when it comes to tradeoffs among

personnel, manpower, and training, all seeking the best return for the investment. Ideally, there should be a balance between defense capability and cost.

- Job proficiency alone does not sum to unit capability. As noted earlier, there are complex interactions among people and job requirements. For example, within work groups the performance of one person is often influenced by the ability and performance of others in the group. Also, individual contributions are partially affected by the availability, type, and sophistication of equipment for getting the job done. Tradeoffs are possible among quality of personnel and sophistication of equipment.
- Not all jobs, independent of how well they are performed, are equally valuable to overall mission capability. For future personnel management systems, the value of both military and private sector jobs must be specified.
- Most performance research to date has focused on first-term airmen, but mission capability depends on the contributions of many people working in concert. It is not necessarily true that the skills needed for success in the first term of enlistment have a substantial bearing on success later in one's career. From an economic standpoint, however, performance in the future is usually regarded as less valuable than performance today. Thus, while PC research should someday consider the broader career progression of personnel, the value of future performance must be discounted for costing exercises.
- Changes in entry standards that affect accession quality mixes will affect the pattern of retention (survival rates). Not everyone who enters stays for a full career. Thus, future value of individual performance must be weighted by survival rates.
- The present value of an individual's military contribution over a career is the sum of each year's contribution, weighted by the associated survival rate and discounted back to the present period. The present values of all individuals in an accession cohort could be summed together, in principle, to measure a Service's expectation of what an accession cohort will contribute over its expected military lifetime. Adjustments to standards would have ripple effects on job performance well beyond the first term, and ultimately affect the lifetime performance of the cohort.
- Different quality personnel have different associated costs for recruiting and training. Typically, higher-quality individuals cost more to recruit because they have more employment or college options. However, high-quality recruits are more likely to succeed in training in shorter times (less total training required and fewer replacements for training failures).

Considering all of these issues, Black suggested that an alternative to building individual performance models would be to focus on the performance of small cohesive groups that produce identifiable products or services. Statistical analyses of group output and the quality of its members avoids the tangle of questions concerning job performance and the interaction of people with each other and with equipment and technology. Research could proceed to minimize the costs of attaining a given level and distribution of group outputs, or to maximize group outputs for a given cost.

Thoughts on PC Implementation

What are the first steps in incorporating PC into manpower planning and personnel standards? One step is to simply facilitate understanding the performance capabilities of the current force. Today's functional managers, commanders, and trainers lack sound, empirical performance data on their personnel. Managers do have knowledgeable insights to the capabilities of forces of given size and quality characteristics, but much of their insight is a product of anecdotal evidence accumulated from personal experience. Analyses of current specialties by grade and skill level using PC as the criterion could be informative and revealing of deficiencies. Such distributions of PC could serve as the basis for more informed judgments about how the force should be shaped. Job redesign and revised standards could achieve desired distributions. Skills required for entry level jobs surely differ from those required at more complex levels. Based on analyses of performance distributions at various levels, standards could be set at several levels to achieve a mix of performance capabilities within an accession cohort. In this way, an entering cohort could be optimally shaped to sustain various career progression characteristics as the cohort ages. In a similar way, manpower standards could be expanded to include different levels of experience desired for various positions (in addition to the basic skill and grade requirements currently specified). Personnel planners would then have the option of allocating personnel on the basis of quality (aptitude) or experience, depending on the characteristic of the available pool of airmen.

Such a shift in manpower and personnel management will not happen at once. For such distribution-based approaches to work, the relationship between criteria, such as PC, and predictors, such as ASVAB scores, must be clearly established. Functional managers must then be introduced to the notion that analyses based on PC can provide useful information in addition to other more familiar sources.

One potential first step toward securing functional manager support would be to solicit their judgments about minimum times to perform job tasks under peak load conditions. The relationship between PC and aptitude scores could be used to set standards to assure peak performance. This approach shifts the focus from changing (and implicitly criticizing) current force characteristics to posturing for future peak load conditions (something functional managers only think about).

Another perspective involves the use of PC in lieu of training performance as the criterion for standards setting. This notion is based on the ability of subject-matter experts (or functional managers) to relate to time-based performance requirements more than quality metrics. Experts would set task benchmark times, such as fastest, slowest, and normal times for acceptable performance. If the accuracy of benchmarks can be consistently confirmed by comparing actual timed measures of tasks with expert estimates, aggregation of benchmarks across tasks should be justified to provide a sense of fastest and normal times required for whole sets of tasks. For desired levels of PC (such as in wartime and peacetime operations), tables of aptitude and experience curves can be constructed to set minimum aptitude required. The

same tables could be used to set multiple cutoffs to achieve specified distributions of PC as discussed earlier.

Another immediate application of PC technology might be to calibrate relative measures of occupational learning difficulty across tasks. Alley (1988) discussed the use of occupational analysis data as criteria for setting enlistment standards; the notion being that specialties can be rank ordered according to difficulty, with higher quality standards justified for more difficult jobs (assuming the jobs are also important). But relative difficulty is only an ordinal-level scale that does not specify precise predictor cutoff scores. PC by definition is a ratio-level construct. For example, if one person does a task in 60 minutes and another does it in 30 minutes, the second person has twice the PC as the first on that task. If functional managers can specify desirable performance times for groups of tasks, the relationship between PC and aptitude can be used to set a minimum quality standard. Aggregation of times across groups of tasks can lead to standards for a specialty. By going through this process with PC-based standards on a number of representative specialties, researchers should be able to establish relationships with learning difficulty that can serve as approximations for standards based on difficulty alone when PC data are not available. Difficulty ratings can be collected operationally along with other task-level information as a matter of course in the occupational survey program. Further research may suggest better measures, such as job incumbent self-reported task times, that could support PC-based standards decisions.

Summary

This discussion of the role of PC in manpower planning and standards setting has not offered solutions to some of the thorny problems facing researchers. It has, however, attempted to offer some ideas to guide the next steps in research. Of most immediate concern is better understanding of the construct itself. Issues such as aggregation of task-level performance requirements within specialties, or specification of acceptable minimum performance needs are dependent on reliable measurement of PC in the first place.

REFERENCES

- Ackerman, P. L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence*, 10, 109-139.
- Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin*, 102, 3-27.
- Ackerman, P. L., & Humphreys, L. H. (1990). Individual differences theory in industrial and organizational psychology. In M. D. Dunnette and L. Hough (Eds.), *Handbook of industrial and organizational psychology*, Volume 1 (2nd Edition). Palo Alto, CA: Consulting Psychologists Press, Inc.
- Alley, W. E. (1988). Occupational learning difficulty. In B. F. Green, Jr., H. Wing, & A. K. Wigdor (Eds.), *Linking military enlistment standards to job performance: Report of a workshop* (pp. 205-213). Washington, D. C.: National Academy Press.
- Alley, W. E., & Teachout, M. S. (1990). *Aptitude and experience trade-offs on job performance*. Paper presented at the annual meeting of the American Psychological Association, Boston, MA.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (ed.), *Educational measurement* (pp. 508-600). Washington, D. C.: American Council of Education.
- Black, M. (1988). Job performance and military enlistment standards. In B. F. Green, Jr., H. Wing, & A. K. Wigdor (Eds.), *Linking military enlistment standards to job performance: Report of a workshop* (pp. 171-197). Washington, D. C.: National Academy Press.
- Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 271-326). Palo Alto, CA: Consulting Psychologists Press, Inc.
- Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt, W. C. Borman, & Associates (Eds.), *Personnel selection in organizations*. San Francisco: Jossey-Bass.
- Borman, W. C., Motowidlo, S. J., & Hanser, L. M. (1983). A model of individual performance effectiveness: Thoughts about expanding the criterion space. In N. K.
- Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology*, 76, 863-872.

- Boyle, E. (1990). *LCOM explained* (AFHRL-TP-90-58). Brooks Air Force Base, TX: Logistics and Human Factors Division, Air Force Human Resources Laboratory.
- Brief, A. P., & Motowidlo, S. J. (1986). Prosocial organizational behaviors. *Academy of Management Review*, 11, 710-725.
- Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, pp. 687-732). Palo Alto, CA: Consulting Psychologists Press, Inc.
- Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection* (pp. 35-70). San Francisco: Jossey-Bass.
- Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology*, 43(2), 313-333.
- Carpenter, J. B., Giorgia, M. J., & McFarland, B. P. (1975). *Comparative analysis of the relative validity for subjective time rating scales* (AFHRL-TR-75-63). Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory.
- Carpenter, M. A., Monaco, S. J., O'Mara, F. E., & Teachout, M. S. (1989). *Time to job proficiency: A preliminary investigation of the effects of aptitude and experience on productive capacity* (AFHRL-TP-88-17). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Carroll, S. J. & Taylor, W. H. (1969). Validity of estimates by clerical personnel of job time proportions. *Journal of Applied Psychology*, 53, 164-166.
- Christal, R. E., & Weismuller, J. J. (1988). Job-task inventory analysis. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government* (Volume II). New York: John Wiley.
- Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harper & Row.
- Demetriades, E. T., & Skinner, J. (1992). *Using subject matter experts to benchmark task performance time scales*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

- DuBois, D., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white-black differences. *Journal of Applied Psychology*, 78, 205-211.
- Ebel, R. L. (1972). *Essentials of educational measurement* (2nd Ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Faneuff, R. S. (1993). *Predicting the productive capacity of Air Force aerospace ground equipment personnel using aptitude and experience measures* (AFIT-GOT-ENS-93M-05). Wright-Patterson AFB, OH: Air University, Air Force Institute of Technology.
- Faneuff, R. S., Valentine, L. D., Stone, B. M., Curry, G. L., & Hageman, D. C. (1990). *Extending the time to proficiency model for simultaneous application to multiple jobs* (AFHRL-TP-90-42). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Belmont, CA: Brooks/Cole.
- Fleishman, E. A., & Mumford, M. D. (1989). Individual attributes and training performance. In I. L. Goldstein (Ed.), *Training and development in work organizations* (pp. 121-182). San Francisco: Jossey-Bass.
- Fleishman, E. A., & Parker, J. R. (1962). Factors in the retention and relearning of perceptual-motor skill. *Journal of Experimental Psychology*, 64, 215-226.
- Frederickson, C. G. (1988). Temporal experience: A two component model. *Perceptual and Motor Skills*, 66, 63-68.
- Grobman, J. H., Quick, D. M., & Weaver, W. M. (1994). *The queuing manpower model (QMAN)*. Unpublished manuscript. Brooks AFB, TX: Manpower and Personnel Research Division, Armstrong Laboratory, Human Resources Directorate.
- Hartley, C., Brecht, M., Pagerey, P., Weeks, G., Chapanis, A., & Hoecker, D. (1977). Subjective time estimates of work tasks by office workers. *Journal of Occupational Psychology*, 50, 23-36.
- Harville, D. L., & Skinner, J. (1993). *Using supervisor estimates of time to job proficiency to set entry standards*. Unpublished manuscript. Brooks AFB, TX: Manpower and Personnel Research Division, Armstrong Laboratory, Human Resources Directorate.
- Hedge, J. W., & Teachout, M. S. (1986). *Job performance measurement: A systematic program of research and development* (AFHRL-TP-86-37). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

- Hedge, J. W., & Teachout, M. S. (1992). An interview approach to work sample criterion measurement. *Journal of Applied Psychology*, 77, 453-461.
- Hogan, H. W. (1978). A theoretical reconciliation of competing views of time perception. *American Journal of Psychology*, 91, 417-428.
- Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisory ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory*. Hillsdale, NJ: Erlbaum.
- Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and applications. *Educational Evaluation and Policy Analysis*, 4, 461-475.
- Jaeger, R. M. & McNulty, S. (1986). *Procedures for eliciting and using judgments of the value of observed behaviors on military job performance tests*. Prepared for the Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council/National Academy of Sciences.
- Lane, N. E. (1987). *Skill acquisition rates and patterns: Issues and training implications*. New York: Springer-Verlag.
- Laue, F. J., Hedge, J. W., Wall, M. L., Pedersen, L. A., & Bentley, B. A. (1992). *Job performance measurement system development process* (AL-TR-1992-0120). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Leighton, D. L., Kageff, L. L., Mosher, G. P., Gribben, M. A., Faneuff, R. S., Demetriades, E. T., & Skinner, M. J. (1992). *Measurement of productive capacity: A methodology for Air Force enlisted specialties* (AL-TP-1992-0029). Brooks AFB, TX: Armstrong Laboratory Human Resources Directorate, Manpower and Personnel Research Division.
- Livingston, S. A., & Zieky, M. (1982). *Passing scores: A manual for setting standards of the performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.
- Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance. *Psychological Bulletin*, 90, 125-152.
- Motowidlo, S. J., & Van Scotter, J. R. (in press). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology*.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3-19.

- Norcini, J. J., Lipner, R. S., & Langdon, L. O. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement*, 24, 56-64.
- Office of the Assistant Secretary of Defense -- Manpower, Reserve Affairs, and Logistics (1980). *Aptitude Testing of Recruits*. Report to the House Committee on Armed Services. Washington, D.C.: U.S. Department of Defense.
- Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Lexington Books.
- Ornstein, R. E. (1970). *On the experience of time*. Baltimore: Penguin Books.
- Poggio, J. P. (1984). *Practical considerations when setting test standards: A look at the process used in Kansas*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document 249267).
- Priestly, J. B. (1968). *Man and Time*. New York: Dell.
- Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, 73(3), 482-486.
- Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, 71, 432-439.
- Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1988). Joint relation of experience and ability with job performance: Test of three hypotheses. *Journal of Applied Psychology*, 73, 46-57.
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, 127-190.
- Skinner, J., Faneuff, R. S., & Demetriades, E. T. (1991). Developing benchmarks to scale task performance times. *Proceedings of the 33rd annual conference of the Military Testing Association* (pp. 399-404). San Antonio, TX.
- Troutwine, R. (1984). Perceived task quality and estimated interval: Foundation for temporal attributions. *Perceptual and Motor Skills*, 58, 100-102.
- Turney, J. R., & Cohen, S. L. (1978). *Perceived work effort as time devoted to an activity* (TP 337) U.S. Army Research Institute for the Behavioral & Social Sciences, Alexandria, VA.

- Wilson, M. A., & Harvey, R. J. (1990). The value of relative-time-spent ratings in task-oriented job analysis. *Journal of Business and Psychology*, 4, 453-461.
- Zakay, D., Lomranz, J., & Kaziniz, M. (1984). Extraversion-introversion and time perception. *Personality and Individual Differences*, 5, 237-239.
- Zieky, M. L., & Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests*. Princeton, NJ: Educational Testing Service.